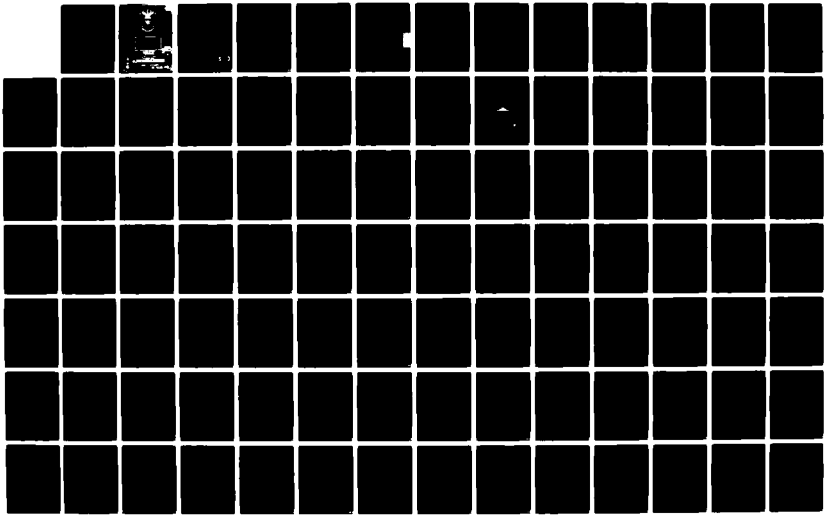
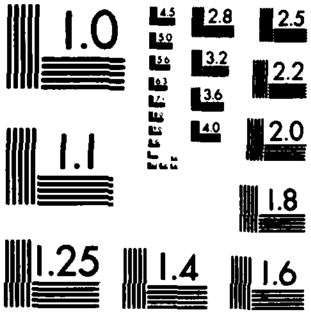
AD-A128·584 EFFECT OF TEST RESULT UNCERTAINTY ON THE PERFORMANCE OF 1/2
A CONTEXT-FREE TR..(U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH SCHOOL OF SYST.. H A BARAN
UNCLASSIFIED 17 DEC 82 AFIT-LSSR-86-82 F/G 15/5 NL




1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

EFFECT OF TEST RESULT UNCERTAINTY
ON THE PERFORMANCE OF A CONTEXT-FREE
TROUBLESHOOTING TASK

Harry A. Baran, GS-12

LSSR 86-82

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deleterious information are contained therein. Furthermore, the views expressed in the document are those of the author(s) and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the Air Training Command, the United States Air Force, or the Department of Defense.

Accession For	
NTIS GRA&I	☑
DTIC TAB	□
Unannounced	□
Justification______	
By______	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

AFIT Control Number LSSR 86-82

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. Please return completed questionnaires to: AFIT/LSH, Wright-Patterson AFB, Ohio 45433.

1. Did this research contribute to a current Air Force project?

a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?

a. Yes b. No

3. The benefits of AFIT research can often be expressed by the equivalent value that your agency received by virtue of AFIT performing the research. Can you estimate what this research would have cost if it had been accomplished under contract or if it had been done in-house in terms of manpower and/or dollars?

a. Man-years ________ $ ________ (Contract).

b. Man-years ________ $ ________ (In-house).

4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3 above), what is your estimate of its significance?

a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments:

________________________ ________________________
Name and Grade Position

________________________ ________________________
Organization Location

FOLD DOWN ON OUTSIDE - SEAL WITH TAPE

AFIT/ LSH
WRIGHT-PATTERSON AFB OH 45433

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE. $300

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 73236 WASHINGTON D.C.

POSTAGE WILL BE PAID BY ADDRESSEE

AFIT/ DAA
Wright-Patterson AFB OH 45433

FOLD IN

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER LSSR 86-82	2. GOVT ACCESSION NO. AD-A128584	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EFFECT OF TEST RESULT UNCERTAINTY ON THE PERFORMANCE OF A CONTEXT-FREE TROUBLESHOOTING TASK		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Harry A. Baran, GS-12		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Systems and Logistics Air Force Institute of Technology, WPAFB OH		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of Communication and Humanities AFIT/LSH, WPAFB OH 45433		12. REPORT DATE 17 December 1982 13. NUMBER OF PAGES 89
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

APPROVED FOR PUBLIC RELEASE: IAW AFR 190-17

18. SUPPLEMENTARY NOTES

24 april 1983

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

TROUBLESHOOTING
MAINTENANCE
ERROR
FAULT DETECTION
DIAGNOSTIC ERROR
HUMAN PERFORMANCE
MALFUNCTION TESTING

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Thesis Advisor: Mr. Russell M. Genet, GM-14

A major problem affecting weapon system support cost and operational capability is the 20-40% troubleshooting error rate experienced by the Air Force and the Navy for all aircraft corrective maintenance actions. Most fault detection problems resemble the problem of signal detection in a background of noise, or in the case of troubleshooting, test result uncertainty. This thesis attempted to determine whether human bias exists under conditions of test result uncertainty such that troubleshooting performance is systematically affected. A knowledge of such bias might be useful in assessing the utility of powerful signal detection-in-noise analytical tools, e.g., the Relative Operating Characteristic (ROC) curve analysis, to improve predictions of troubleshooting performance, and reduce troubleshooting error by allowing man-machine troubleshooting systems to be optimized on the basis of the response characteristics of both machine and man. The experiment consisted of sixty-four subjects performing a simulated electronics troubleshooting task at four levels of test result error; zero error, 25% error, 50% good called bad error, and 50% bad called good error. Results did not indicate the existence of significant differences in troubleshooting performance under the latter three treatment conditions; only between the first and any of the latter three.

LSSR 86-82

EFFECT OF TEST RESULT UNCERTAINTY ON THE PERFORMANCE OF A CONTEXT-FREE TROUBLESHOOTING TASK

A Thesis

Presented to the Faculty of the School of Systems and Logistics
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the Requirement for the
Degree of Master of Science in Logistics Management

By

Harry A. Baran, B.A.
GS-12

December 1982

This thesis, written by

Mr. Harry A. Baran

has been accepted by the undersigned on behalf of the faculty of the School of Systems and Logistics in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN LOGISTICS MANAGEMENT
(ACQUISITION LOGISTICS MAJOR)

DATE: 17 December 1982

Russell Mc

COMMITTEE CHAIRMAN

FACULTY READER

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Air Force Human Resources Laboratory (AFHRL) for the opportunity to attend AFIT. Specific thanks are extended to Colonel Donald C. Tetmeyer for encouraging me to enter the program; to Messrs Edward Boyle, Frederick Davis, Terry Miller, and Jeromy Notestine for their help in performing the thesis experiment; and to Ms Elizabeth J. Allebach for her invaluable aid in typing and compiling this document, and the graciousness with which it was given.

Lieutenant Colonel Ronald G. Blackledge deserves thanks for the guidance he afforded me in his capacity as thesis faculty reader, instructor, and friend. Mr Russell M. Genet also deserves thanks for agreeing to act as thesis advisor.

My most profound thanks are extended to my wife, Kazue, and to my son, Harry. Their forebearance from obloquies concerning my obsequious acceptance of the baneful and vicissitudinal character of this thesis experience can never be understood, but will always be appreciated by me. The most meaningful aspect of that experience was an empirical knowledge of the extent of their love and devotion in the face of my forfeiture to this thesis of time best spent in attending to the pleasant duties of marriage and parenthood.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure | Page

CHAPTER I

INTRODUCTION

BACKGROUND

Air Force weapon systems are becoming increasingly complex. A major impact of this is a high error rate in the troubleshooting of systems to detect and isolate faults. Gibson (1982) has cited two types of error as being major problems in maintaining Air Force systems: Type I, the removal of good system components for repair and; Type II, the failure to detect malfunctioning components which should be removed from the system.

Gibson (1982) and Lipa (1982) reported studies which indicated that the Type I maintenance error, known as a retest ok or "RTOK", occurs at an average rate of thirty percent for all maintenance actions involving avionics line replaceable units (LRUs). Orlansky and String (1981) referenced seven studies which found that non-faulty components are removed from Navy aircraft in up to forty three percent of all corrective maintenance actions, and that such removals account for up to thirty two percent of all maintenance manhours. These findings were supported by the membership of the Built-In-Test Equipment Workshop (IDA,1981:S1-10) which estimated the current Air Force rate of unnecessary removals at between twenty and forty percent of all corrective maintenance actions.

Orlansky and String (1981:8-9) reported that technicians commit a Type II error, i.e., fail to find a faulty part, or damage a good part in about ten percent of all corrective maintenance actions. The extent of this type of error has been much more difficult to confirm than that of Type I. However, it is widely acknowledged as being a frequently occurring problem, especially with respect to highly sophisticated aircraft such as the Air Force's F-15 and the Navy's F-14 (Gibson,1982; Lipa,1982).

The impact of diagnostic errors is felt at nearly all levels of operations and support.

> "Some of these errors can produce significant effects, e.g., abort an operation, require repetition of the troubleshooting and repair actions, waste spare parts, place an additional load on the maintenance activity, or perhaps lead to an injury or accident [Orlansky and String,1981:4]."

Their impact results in increased support cost and reduced operational capability. This effect was confirmed by Clyman, Grentz, and Schultz who stated that:

> "Unnecessary unscheduled maintenance actions contribute a large share to the operation and support cost of airborne weapon systems . . . which represents a continually increasing proportion of the defense budget [1978:iii]."

They concluded that the incidence of maintenance troubleshooting errors is critical to the fulfillment of Air Force mission requirements both in terms of cost and performance.

Advances in the design of weapon systems have been accompanied by new hardware and software developments to improve fault detection and fault isolation, e.g., automated built-in-test (BIT) devices. However, " . . . there does not appear to be sufficient emphasis on testing systems from a human reliability standpoint [Comptroller General of the United States,1981:277]." Total diagnostic system performance capability requirements include, and will in the foreseeable future, the contribution of the human operator working on the basis of test result information and logic (IDA,1981:S6). The Built-In-Test Equipment Workshop sponsored by the Institute of Defense Analyses (IDA,1981) concluded: (1) that current practices in specifying the diagnostic capability.to be associated with a weapon system do not sufficiently address the human operator's contribution (including his biases) to fault detection and isolation, and (2) that research is required to define approaches to specify, predict, and evaluate the incidence of error within system diagnostic activities.

According to Gold, Kleine, Fuchs, Ravo, and Inaba (1980:14), maintenance technicians can produce three kinds of errors in organizational maintenance: replace a good unit, fail to replace a bad unit, or damage the system in some way. Disregarding the effects of induced damage due to troubleshooting, there are four basic outcomes possible in a troubleshooting decision; (1) a malfunction exists and is detected, (2) a malfunction exists and is not detected, (3) no malfunction exists and none is detected, and (4) no malfuncion exists but one is erroneously detected. These possible outcomes are depicted in Table 1 as a truth table.

TABLE 1

TROUBLESHOOTING DECISION TRUTH TABLE

	Malfunction Present	Malfunction Not Present
Malfunction Detected	CORRECT DECISION	TYPE I ERROR (False Alarm)
Malfunction Not Detected	TYPE II ERROR (Missed Bad)	CORRECT DECISION

As can be seen in Table 1, there are two types of errors possible. Type I (False Alarm), if committed, might result in the unnecessary removal of equipment. Type II (Missed Bad), if committed, might result in faulty equipment being allowed to remain in the system.

> "Most diagnostic problems consist of a sequence of two-alternative decisions: the generic case being the detection of a signal in a background of random interference or 'noise' [Swets and Pickett, 1980:107]."

In most maintenance troubleshooting situations involving complex systems, clear indications of system component failure are rare. The role of the troubleshooter can be characterized as being interactive with, rather than as simply responsive to, test equipment results (Rouse,1978; Swets and Pickett,1980). Given the existence of noise or ambiguity in troubleshooting test results for complex systems[1] and the resemblance of diagnostic decisions to those which are common to the signal-embedded-in-noise class of problems, the possibility exists that operator bias may be a factor which acts to selectively inhibit or facilitate fault detection in a manner similar to that which has been demonstrated for signal-in-noise detection. Thus, analytical techniques proven useful for the signal-in-noise detection class of problems may be applicable to that of troubleshooting fault detection.

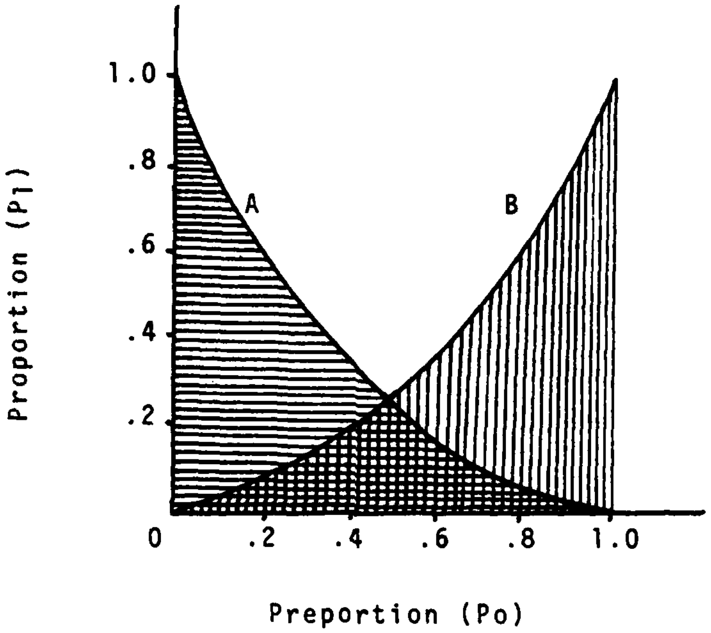
The effect of noise on signal detection has been the topic of a large body of research dating back to the early 1950's. One of the major development stemming from

[1] This was a finding of the Built-In-Test Equipment Workshop (IDA,1982:5-8).

that work is the Relative (or Receiver) Operating Characteristic (ROC) curve shown in Figure 1.

FIGURE 1

RELATIVE OPERATING CHARACTERISTIC CURVE



Curve A - Proportion of True-Positive to True-Negative Detections[2]
Curve B - Proportion of False-Positive to False-Negative Detections[2]

▤ Type I Error

▥ Type II Error

▦ Total Error

[2] Refer to Table 1.

Developed by Peterson, Birdsall, and Fox (1954), the ROC curve is still a primary means of analysis for decision problems entailed in signal detection tasks. Its development was based on statistical developments by Neyman and Pearson (1933) and paralleled certain developments by Wald (1950), who had earlier demonstrated how decision problems are amenable to probabilistic and statistical analysis. Its use and the application of the analytical techniques of statistical decision theory and signal detection theory to human discrimination and decision making activities has been described in text book form by Green and Swets (1966).

The ROC curve is a means of analyzing the various tradeoffs among proportions of correct detections, Type I, and Type II errors, as decision criteria are systematically varied. Its chief use is to account for the effect of operator bias on signal detection performance for predictive purposes. According to Green and Swets (1966:240), it has been applied extensively in perceptual and cognitive psychological studies as a predictor of operator performance which takes into account individual differences and the effect of incentives which may be associated with performance. It has also been reported by Swets, et al (1979) and Swets and Pickett (1980:28-51), to have been successfully applied in the field of medical diagnostic decision making. In addition, Kerr (1976) has described its use in tests of inertial navigation systems. In that application, the ROC relationships were used to compensate for the influence on decision making exerted by a human's ability to recognize trends occurring over time in test results.

The above successful applications of the ROC curve analysis tend to support the notion of its application to reduce maintenance troubleshooting error. An example of such an application would be in the establishment of response cut-off scores or threshold values in system test equipment. Such threshold values are decision criteria parameter value limits beyond which test equipment will indicate an out-of-tolerance or fault condition. The purpose would be to go beyond current practices to include the response characteristics (bias) of the human operator as a determinant of the response characteristics of the test equipment. The overall objective would be to treat the troubleshooting function as a man-machine system based on the expectation that an optimization of total system response thresholds would lead to a reduction of troubleshooting error.

Current practice in setting test equipment response thresholds does not specifically consider the effect on troubleshooting performance of operator bias or motivation, but does tend to favor the incidence of Type I (False Alarm) errors. The original reasoning behind this posture was that the resource expenditures which result, although recognized as considerable, were thought to be of secondary importance to ensuring a capability to fulfill mission objectives. However, experience (see Gibson,1982; Lipa, 1982; Orlansky and String,1981; IDA,1981) has shown that the expenditures associated with Type I errors have grown far beyond what may have been originally anticipated. Whether to maintain a posture which favors the commission of Type I errors is, at present, a subjectively based policy decision which could be substantially aided by a more precise knowledge concerning the existence and direction of human bias, i.e., a predisposition to perform better or worse under various conditions of reward or of test equipment response threshold setting. It could be aided even further by the use of a performance predictor which could effectively control for any effects of that nature.

Descriptions of the ROC curve and the history of its development (Egan,1975; Swets,1973; McNichol,1972; Green and Swets,1966; Swets,1961) indicate that the relationships depicted in the ROC curve are predictors of performance which are not biased by operator effects. In fact, their most common use is to describe a baseline level of performance upon which the effects of bias can be superimposed to clearly indicate performance differences directly attributable to operator bias. The most unique characteristic of an ROC curve analysis is its ability to describe performance baselines and the effect of bias as separate entities.

If operator bias is a significant factor in determining troubleshooting performance, than use of the ROC approach to its prediction could be an effective means to reduce troubleshooting error. For example, if it could be demonstrated that operators perform better or worse under certain conditions of noise (test result uncertainty), e.g., test equipment response threshold settings, the ROC curve analysis would be shown to be a potentially effective tool to describe pertinent performance vs test equipment bias relationships and capitalize on the operator bias effects to improve performance. Stated another way, the identification of a systematic operator bias effect on task performance is a prerequisite first step in determining whether an effort should be made to investigate the potential of the ROC curve analysis for improving maintenance troubleshooting performance.

Despite the strong similarity between the typical troubleshooting situation and the classical signal-detection-in-noise situation, there has been little research to document either the nature or the effect of external (equipment bias) or internal (human bias) noise in a troubleshooting context. With the exception of Pieper and Folley

(1967) who examined the effect of withholding test result information on troubleshooting accuracy and time requirements, the effect of noise on the human operator's propensity to compensate for or optimize test result information has not yet been investigated. In addition, Orlansky and String state that " . . . surprisingly little objective data are available to document how well maintenance technicians do what they are supposed to do [1981:1]."

Maintenance operators have indicated their awareness that test results are often ambiguous and contain a noise component (Clyman, Grentz, and Schultz,1978:47). Real world maintenance conditions foster an awareness of the existence of uncertainty in the diagnostic precision of test equipment and the possible occurrence of anomolies, malfunctions, and environmental effects which might affect their output. An experienced maintenance technician might easily become suspicious of test result information and prone to initiate compensatory behavior (Kerr,1976:122).

If a troubleshooter assumes the role of compensator, an action which is often described as being the human's chief contribution in man-machine systems (Lomov, 1979; Rasmussen and Rouse,1981), he may well be a source of non-randomly distributed internal noise or bias. Evidence which supports the notion of a human bias effect on the performance of troubleshooting activities has been reported by Rouse (1978). He found that humans tend to discount the value of information about what has not failed in searching for the source of malfunctions.

In an evaluation of troubleshooting behavior, Swets and Pickett (1980:110) concluded that the troubleshooter or diagnostician wants to make fewer incorrect decisions than correct decisions. He is concerned with the total value of his performance, and so desires to minimize errors weighted by their importance to some overall objective, i.e., to maximize the expected value of a decision or to minimize the maximum risk rather than simply to maximize the percentage of correct decisions. An example of such behavior is the toleration of false alarms (Type I errors) in order to be relatively certain that all faults which could cause mission failure will be detected; exactly the thought process undertaken by personnel responsible for establishing test equipment thresholds.

An answer to the question of whether the operator is a source of systematic troubleshooting performance bias arises as a requirement for research in at least two ways: (1) as a potential source of error in current procedures to predict maintenance troubleshooting performance, and (2) as a potential indicant of whether it would be advisable to investigate the application of a powerful signal detection-in-noise analytical tool to the prediction of troubleshooting performance.

Troubleshooting performance can be measured in terms of the types of error committed, i.e., false alarms or failures to detect faulty system components. The effect of test result uncertainty or noise can be operationalized in an experimental context by altering test result information such that it simulates the effect of test equipment response threshold settings which favor the commission of one or the other type of troubleshooting error. Thus, the means are available to experimentally investigate whether there is an operator bias present in man-machine based troubleshooting tasks which has a systematic effect on the commission of troubleshooting errors.

If the existence of human bias as described above can be experimentally demonstrated, it is important to determine the direction of its effect, i.e., whether the bias effect is selectively facilitative or inhibitive of task performance under conditions of test result uncertainty which favor either the occurrence of troubleshooting errors in which the operator either calls a good component bad (False Alarm) or errors in which he calls a bad component good (Missed Bad).

The following section contains a primary and a follow-up research question which address the need to determine the existence and direction of operator bias in troubleshooting task performance.

RESEARCH QUESTIONS

Research Question One

The primary research question was:

> In the presence of noise, is there a human operator bias present in man-machine based troubleshooting tasks which has a systematic effect on performance as measured by the commission of Type I (False Alarm) and Type II (Missed Bad) errors?

Research Question Two

In the event that the existence of a systematic human operator bias could be demonstrated, it was important to determine its directional impact on task performance.

The second research question was:

> If the existence of a systematic human operator bias in man-machine based troubleshooting tasks can be demonstrated, what is the direction of its impact, i.e.,does the bias selectively inhibit or facilitate task performance under conditions of test result uncertainty which favor either the commission of a Type I or Type II error?

SUMMARY

There is considerable concern in the Air Force about the high incidence of maintenance troubleshooting errors. This concern is growing because the increasing sophistication of equipment in modern weapon systems has been described as both a requirement for future mission success and a reason to expect an increase in troubleshooting errors (Comptroller General of the United States, 1981; Perry,1979,1973). Such errors result in high system support costs and adversely affect a weapon system's operational capability.

The incidence of troubleshooting error is a function of the information used by a troubleshooter to diagnose a system and the manner in which he used it in a decision analysis to identify and localize system faults. Much of that information is obtained from test equipment whose response thresholds for fault indication have been established with little or no consideration of the role of the human operator as an information processor who is subject to both external (equipment bias) and internal (human bias) sources of noise which affect his performance.

The existence of external noise, i.e., test result ambiguity, in the task of troubleshooting, allows for a comparison to be made between that task and the task of signal detection in noise. Similarities between the two tasks support the possibility that analytical tools successfully applied in the prediction of performance for the latter may be applicable in the prediction of performance for the former; specifically, the Relative Operating Characteristic (ROC) curve. The ROC curve may afford an improved means to perform trade-offs between proportions of correct decisions, Type I errors (False Alarms), and Type II errors (Missed Bad), given that internal noise (human bias) has a significant effect on troubleshooting performance. The expected effect of such an improvement is a reduction in Air Force maintenance troubleshooting errors.

Chapter I has (1) provided an explanation of the objective of the research performed in this thesis, and (2) posed two research questions concerning the existence and directionality of human operator bias in the performance of troubleshooting tasks which systematically inhibits or facilitates task performance as a function of test result uncertainty (external noise). Chapter II describes the research approach developed to operationalize and answer the research questions. Chapter III reports the results and findings of the research. Chapter IV presents the conclusions and recommendations drawn from the research.

CHAPTER II

RESEARCH METHODOLOGY

In this chapter, the experimental approach to answer the research questions is developed. Next the experimental task is described. Following that, the experimental design, experimental conditions and controls, and supporting hardware and software are discussed. In the following section the primary research question is stated as a statistically testable hypothesis. The chapter is concluded with a description of the statistical procedures used in testing the research hypotheses.

APPROACH

A laboratory study approach was taken to answer the research questions of this thesis. A field study approach, using field observations of actual maintenance activities and historical maintenance data, was considered but was determined to be inappropriate for the purposes of this thesis. The reasons for this decision included; (1) difficulties which would be encountered in controlling the environment, (2) cost and time constaints, and (3) difficulties entailed in selecting an actual troubleshooting problem and test equipment from which generalizable conclusions could be drawn with respect to other troubleshooting problems and test equipment. Table 2 summarizes the advantages and disadvantages of the field vs the experimental study approach for this research. The experimental task was patterned after one used by Rouse and his associates in the late 1970's (Rouse,1979a,1979b,1978; Hunt and Rouse, 1981; Johnson and Rouse,1982). It simulated the troubleshooting of an electronic system consisting of twenty-five component elements, with a fixed pattern of interconnections and a maximum of one faulty element per problem. It was chosen because it is not specific to any particular Air Force troubleshooting task but is representative of a large variety of fault diagnosis tasks found within Air Force maintenance activities. Such a task is said to be "context-free." Information presented to subjects was generated and controlled by a series of computer programs which also recorded and processed subject responses. Presentations to subjects were made using a cathode ray tube (CRT) display. Subject responses were made using a computer keyboard.

The experiment consisted of five task sessions, three of which were conducted for training purposes to ensure that subjects could perform the experimental task at a baseline level of proficiency. The final two sessions provided experimental data. During each session, subjects

TABLE 2

RESEARCH APPROACH ALTERNATIVES
DECISION TABLE

APPROACH	ADVANTAGES	DISADVANTAGES
FIELD STUDY	(1) High Validity in Terms of Subjects and Troubleshooting Environment (2) Ease of Obtaining Subjects	(1) Expensive (2) Results May Not be Generalizable for All Troubleshooting Tasks or Test Equipment (3) Excessive Time Requirements (4) Difficult to Control Environment
LABORATORY STUDY	(1) Relatively Inexpensive (2) Quickly Accomplished (3) Controlled Environment (4) Easy to Replicate (5) Generalizable Troubleshooting Task	(1) Difficult to Obtain Subjects (2) Low Validity in Terms of Troubleshooting Environment

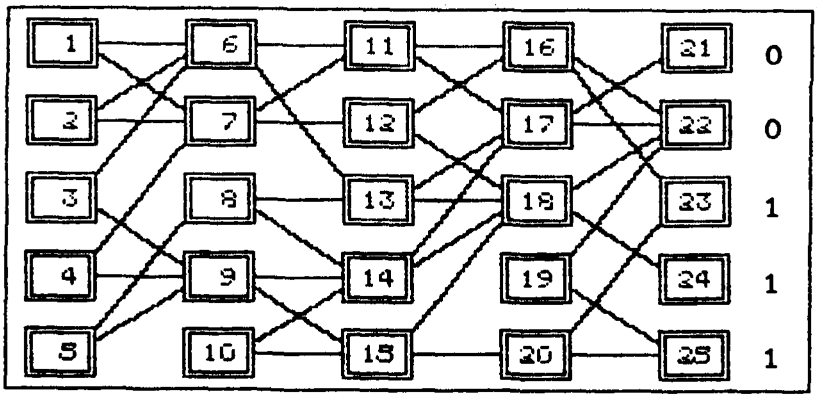
were presented a series of thirty troubleshooting problems to solve. Each of the five problem sets was unique but comparable to the others in terms of difficulty and the total number of faults to be found. Problems containing a fault (approximately fifteen in each set) were randomly distributed within each problem set. In addition, the presentation sequence of problem sets to subjects in the two experimental sessions was also random.

EXPERIMENTAL TASK

Subjects were confronted with a display resembling a simplified electronic system schematic diagram as is shown in Figure 2.

FIGURE 2

TROUBLESHOOTING PROBLEM DISPLAY



A problem began with the display of that network of system components, with Overall Test outputs beside and to the right of the network, as is shown in Figure 2. On the basis of this information, the subjects' task was to perform a series of tests which would culminate in their locating and designating for replacement the system component which had failed.

Fault diagnosis in the situation presented to the subjects involved dealing with a network of dependencies among system components which determined their abilities to produce acceptable output signals. For example, (refer to Figure 2) boxes 21 and 22 are producing unacceptable signals, as is indicated by the zeros shown in the test outputs to the right of them. Since there can be a maximum of one faulty component in a problem (and here there are two indications), it cannot be either box 21 or box 22, but some common box to their left in the network which is passing on a bad signal to both of them. Examination of Figure 2 reveals that there is only one box which feeds signals to and only to boxes 21 and 22. It is box 17, the box which must be faulty.

There were two kinds of tests to determine whether a fault was present in the network, and if so, to localize it. Those tests were; (1) an Overall Test which examined the final output of the system (those from boxes 21, 22, 23, 24, and 25) to determine if the entire system was functioning correctly without fault, and (2) a Localization Test which examined one or several components of the system which were directly connected to each other to determine if there was a fault present. In the latter, the subjects specified which box(es) were to be included in the test. Subjects were informed that the architecture of the system would remain constant for all problems and that there would never be more than one faulty component in the system. After undergoing a training program to ensure basic competency with the equipment and the experimental task, subjects were provided two experimental sessions in which errors were introduced into the information presented to them.

Four levels of test result error were examined within the experimental sessions; (1) the absence of error, (2) twenty-five percent of all "good" and "bad" test result indications made incorrect, (3) fifty percent of the "good" test result indications made to wrongfully read "bad" while all of the "bad" indications remained unchanged (a negative bias), and (4) fifty percent of the "bad" test result indications made to wrongfully read "good" while all of the "good" indications remained unchanged (a positive bias). Error was assigned to individual problems within a given problem session on a random basis in order to simulate the effect of unreliable test equipment.

Subjects were instructed to; (1) first run an Overall Test, (2) then run a Localization Test if the Overall Test indicated the presence of a faulty component and they thought it was necessary, (3) continue to run Localization Tests until the fault had been identified, (4) replace the component they had identified as faulty, run another Overall Test to confirm that the system was operating without fault, and (5) signal completion of the problem. Subjects were also told that the test equipment would lie to them some of the time. During the training sessions, subjects were informed that there were only three kinds of ambiguous Overall Test results, given that the test equipment was not lying to them. Those cases were when the Overall Test indicated that either three, four, or five of the system components in the column farthest to the right in the system network were "bad." Under any of those circumstances, one or more Localization Test would be mandatory. Examples of each kind of test were provided in the training sessions (see Appendix A).

Due to the architecture of the system network and the fact that there can only be one faulty component in a problem, there are only seventeen possible outcomes of an Overall Test. One is where all five components in the column farthest to the right in the network (the only ones tested in an Overall Test because that test examines only the final system outputs) are indicated as being "good." Three others are the test outcomes mentioned in the preceding paragraph. The remaining thirteen possible test outcomes can be traced back to a single component on the basis of the Overall Test and logic.

EXPERIMENTAL DESIGN

The experimental design used in this research was patterned after the classical Completely Randomized Split-Plot Factorial Design (Kirk,1968:298-307). There were three treatments within the experiment; (1) Test Result Error, (2) Subject-To-Experimenter Assignment, and (3) Problem Set Order Of Presentation.

The treatment effect of primary interest was that of Test Result Error. The other two treatments were included in the experiment as means to examine the possibility that there might be a Subject-To-Experimenter Assignment or Problem Set Order Of Presentation interaction effect with the Test Result Error effect. The purpose of having a Subject-To-Experimenter Assignment Treatment in the experiment was to increase the generalizability of results. The purpose of having a Problem Set Order Of Presentation Treatment in the experiment was to control for possible differences between problem sets.

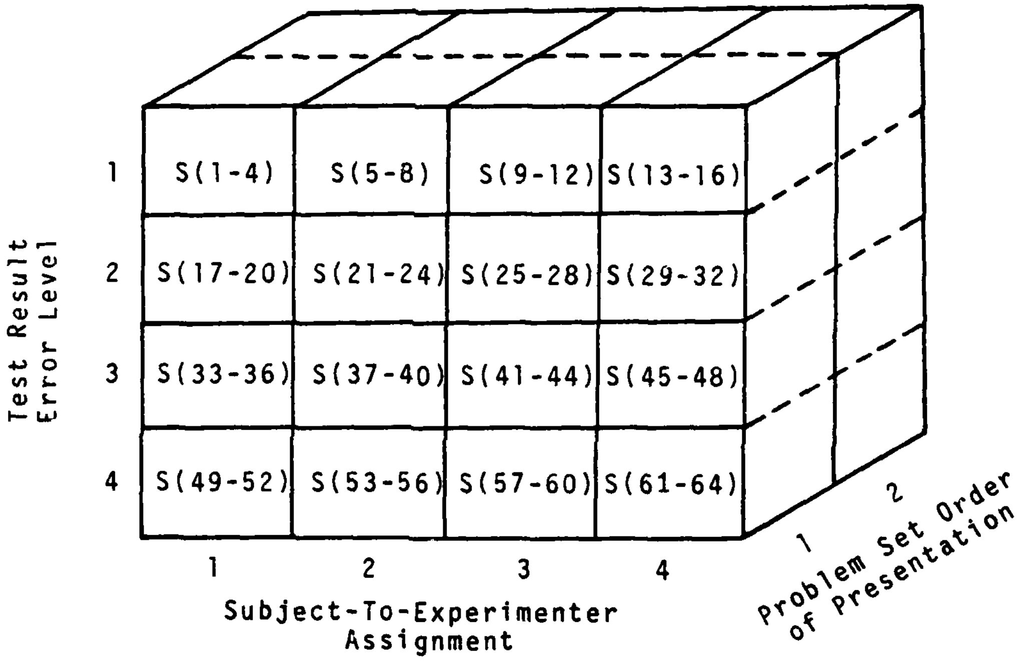
Figure 3 depicts the experimental design and defines the treatments by subjects breakout. As can be seen in Figure 3, there were sixty-four paid subjects, four levels of Test Result Error, two Problem Set Orders of Presentation, and four Subject-To-Experimenter Assignment conditions. The number of experimenters was established as an arbitrary minimum for the timely conduct of the experiment. The number of problem sets was established on the basis of maximum time limits for conducting an experimental session in which subjects' fatigue and interest could be maintained at reasonable levels. The quantity of subjects required in the experiment was established by a calculation of the minimum treatment cell count required to achieve a desired power level for the F statistic test which was used in the analysis of the experimental data, and by the experimental design employed in the experiment.

Independent Variables

There were three independent variables; (1) the level of error introduced into the information provided by the tests performed by subjects within the experimental task, (2) the assignment of a subject to a particular experimenter in the conduct of the training and experimental sessions, (3) the sequence in which the two experimental problem sets were assigned to a subject in the two experimental sessions.

FIGURE 3

BLOCK DIAGRAM OF EXPERIMENTAL DESIGN


Test Result Error Level
Subject-To-Experimenter Assignment
1
2
Problem Set Order of Presentation

	1	2	3	4
1	S(1-4)	S(5-8)	S(9-12)	S(13-16)
2	S(17-20)	S(21-24)	S(25-28)	S(29-32)
3	S(33-36)	S(37-40)	S(41-44)	S(45-48)
4	S(49-52)	S(53-56)	S(57-60)	S(61-64)

S - Subjects under a treatment condition

Test Result Error

There were four levels of Test Result Error:

1. Zero percent of the test results erroneous

2. Twenty-five percent of the test results erroneous

3. Fifty percent of the "good" test results changed to read "bad"

4. Fifty percent of the "bad" test results changed to read "good"

Subject-To-Experimenter Assignment

There were four possible assignments of a subject to an experimenter. Once an assignment was made, all sessions of a given subject were, without exception, conducted in accordance with that assignment. Each experimenter was assigned one quarter of the subjects.

Problem Set Order of Presentation

There were two possible sequences in which the two experimental problem sets could be assigned to a subject in the two experimental sessions.

Dependent Variables

There were three dependent variables in this experiment. Each is a measure of subject performance on the experimental task. The first was the number of False Alarms; "good" components mistakenly replaced. The second was the number of Missed Bad; "bad" components which were not replaced. The third was the Bonus Score earned by a subject on the basis of task performance; a composite score which took into account penalties for incorrect responses and inefficiencies on the part of the subject in completing the task. An explanation of the Bonus Score and its calculation is contained in Appendix A. It was included in the experiment as a basis for additional monetary payment (an incentive for subjects to do their best), and also to provide a baseline quantification of subject motivation to perform the experimental task.

All of the dependent variables were continuous and measurable as ratio scalar quantities. The selection of dependent variables was predicated on the results of a comprehensive investigation of performance measures reported by Henneman (1981) and Henniman and Rouse (1982) which produced a set of twenty candidate measures which appeared to be appropriate for troubleshooting tasks. Rouse and Hunt (1982) examined these twenty measures within the context of two experiments of a nature similar to that of the experiment performed in this thesis. Results were unequivocal on the basis of correlation, regression, and factor analysis; the only unique dimensions were error, efficiency, and time.

EXPERIMENTAL CONDITIONS AND CONTROLS

The following subsections describe conditions and controls which were made part of the experiment to (1) increase the generalizability of results across systems and troubleshooting tasks, (2) increase the generalizability of experimental results on the basis of their freedom from any confounding effects attributable to the experimental environment, and (3) reduce experimental error attributable to subject variability. They are presented below as they relate to a particular facet of the experiment.

Experimental Presentation

The experimental task was context-free in terms of its lack of specificity to real world systems and troubleshooting tasks. The selection of a context-free task for use in this experiment was predicated on the findings of Hunt and Rouse (1981) and Johnson and Rouse (1980), who experimentally demonstrated that subjects' troubleshooting performance using a context-free simulation was highly correlated with their performance in troubleshooting real equipment. The use of a context-free task in this experiment provided a basis for generalizing results to a wide variety of real systems and troubleshooting tasks.

Experimental Subjects

Sixty-four paid subjects were selected from a population of college students. The main reason for doing so was the nonavailability of maintenance troubleshooting personnel for participation in the experiment. However, the use of paid subjects who were relatively naive in terms of the experimental tasks, although less desirable than the use of actual Air Force maintenance personnel, obviated much of the learning or experience factor to which the existence of human bias has often been attributed. It was thought that, in the event that a bias effect could be demonstrated for subjects not trained or experienced in real world troubleshooting activities, it may well have been even more in evidence if experienced maintenance personnel were to have been used as subjects.

A control was established for subject training. Three task sessions were devoted to ensure that each subject achieved a baseline level of task performance under conditions of zero Test Result Error.

Subject-To-Experimenter Assignment

Because of the large number of subjects in the experiment and the time required to run five sessions for each subject, four experimenters were used. Although this required that extra controls be added to the experiment, it also served to increase the generalizability of experimental results. Three controls were established to address the possibility of experimental error due to experimenter variability; (1) subjects were sequestered during the experimental sessions, (2) experimenters used a highly proceduralized standardized protocol for the administration of training, and (3) subjects were randomly assigned to experimenters.

Problem Set Comparability and Ordering

The design of the experiment included a replication of the experimental task session for all subjects. To avoid learning effects, two unique problem sets were administered to each subject. They were devised to be very similar in difficulty and content. Three controls were associated with the problem sets; (1) problems containing faults were randomly distributed within a problem set, (2) the order of problem set presentation to subjects was made to be random, and (3) scores for all measures of task performance were normalized to compensate for any differences between problem sets.

SUPPORTING HARDWARE AND SOFTWARE

Three 48K Apple II-Plus microcomputers were used in the experiment. Associated peripherals consisted of three cathode ray tube (CRT) displays, two printers, four floppy disk drive units, and a modem. They were used to develop the computer programs used in the experiment, present the experimental tasks to the subjects, and record and process experimental data. Analysis of the data was performed using the ASD Computer Facility's CDC 6600 computer located at Wright-Patterson AFB, along with the BMDP2V statistical analysis program package.[3]

Several computer programs were developed to generate experimental displays, provide for a discourse between subjects and the computer during training and experimental sessions, create and maintain files for recording and processing subjects' responses and other data concerning the logistics of running each of sixty-four subjects through three training and two experimental sessions. Short descriptions of these programs are provided in Appendix B.

RESEARCH HYPOTHESES

The primary research question can be operationally stated in terms of a statistically testable hypothesis as follows:

> H_0: The group means of scores on the False Alarm and Missed Bad measures of task performance were each not significantly different under Level 3 (50% of the "good" test results changed to read "bad") and under Level 4 (50% of the "bad" test results changed to read "good") of the experimental treatment of Test Result Error.
>
> H_a: The difference between the means was significantly different.

[3] The BMDP2V is a statistical analysis program package made available by the Health Sciences Computing Facility of the University of California, Los Angeles.

If the null hypothesis is not rejected, the follow-up research question is moot. If the null hypothesis is rejected, the secondary research question will then be answered on the basis of a determination of the directionality of the significant differences found among the group means of the task performance measure scores. That determination is to be made by examining the means of those scores.

The primary research hypothesis was the subject of an a priori analysis of the experimental data, the plans for which will be described in the next section of this chapter. That analysis also addressed a second aspect of the primary research question which was operationalized in terms of a the follow-up research hypothesis. Both the primary and secondary research hypotheses were broken down into four sets of research hypotheses to accommodate their separate testing in terms of both False Alarm and Missed Bad task performance measure scores. Testing of the follow-up research hypothesis was not conditioned upon the outcome of the test of the primary research hypothesis. Each of the four sets of hypotheses are stated in the next section of this chapter.

DATA ANALYSIS PLAN

A Priori Analysis

The a priori analysis of the experimental data was comprised of a set of two a priori orthogonal comparisons of group means for each of two (False Alarm and Missed Bad) of the three measures of troubleshooting performance. The planned comparisons were accomplished prior to performing an overall analysis of the entire set of experimental data. They are described below.

The task performance measure False Alarm was the subject of a comparison between Level 3 (50% of the "good" test results changed to read "bad") and Level 4 (50% of the "bad" test results changed to read "good") of the Test Result Error treatment variable, and also between Level 2 (25% of the test results erroneous) and Levels 3 and 4 combined of that same treatment variable. The same two comparisons were also made for the task performance measure of Missed Bad. It was expected that task performance, as measured by subjects' scores for False Alarm and Missed Bad (which are measures of signal-detection-in-noise problem Type I and Type II errors, respectively)[4], would be

[4] A Type I error is said to be committed in a signal-detection-in-noise problem if a signal is detected when none is present. A Type II error, for that class of problems, is a failure to detect a signal when one is present.

affected differently by Level 3 (50% of the "good" test results changed to read "bad") and Level 4 (50% of the "bad" test results changed to read "good") of the Test Result Error treatment variable. The comparisons between group means for Level 2 (25% of the test results erroneous) and those for a combination of Level 3 (50% of the "good" test results made to read "bad") and Level 4 (50% of the "bad" test results made to read "good") of that treatment variable were made on the basis of an expectation that non-directional test result errors (typified by the kind of errors introduced under Level 2 of the Test Result Error treatment variable) would affect task performance differently than directional test result errors (typified by those introduced under Levels 3 and 4 of the Test Result Error treatment variable). The a priori analysis addressed four sets of hypotheses:

Hypothesis Set 1

H_0: The mean of scores for the task performance measure False Alarm under the treatment condition of Test Result Error Level 3 (50% of the "good" test results changed to read "bad") was not significantly different from the mean of scores for that measure under the treatment condition of Test Result Error Level 4 (50% of the "bad" test results changed to read "good").

H_a: The difference between the means was significantly different.

Hypothesis Set 2

This hypothesis set is the same as the preceding set except that the task performance measure Missed Bad was examined instead of the task performance measure False Alarm.

Hypothesis Set 3

H_0: The mean of scores for the task performance measure False Alarm under the treatment condition of Test Result Error Level 2 (25% of the test results erroneous) was not significantly different from the mean of scores for that measure under the treatment condition of Test Result Error Level 3 (50% of the "good" test results changed to read "bad") and Level 4 (50% of the "bad" test results changed to read "good") combined.

H_a: The difference between the means was significantly different.

Hypothesis Set 4

This hypothesis set is the same as the preceding set except that the task performance measure Missed Bad was examined instead of the task performance measure False Alarm.

The statistical procedure used in the a priori orthogonal comparisons of the paired group means was the Student's T-Test analysis for paired comparisons. As was stated in the previous section of this chapter, the primary research hypothesis was tested within the a priori analysis.

Overall Analysis

An overall analysis of the entire set of experimental data was conducted to identify significant effects among all of the variables in the experiment, examined as a single group. The statistical procedure used in an overall test of the significance of differences between the group means of task performance measure scores, under all experimental treatment conditions, was the Analysis of Variance (ANOVA). Results were used as criteria for deciding whether an a posteriori analysis should be performed on the experimental data, i.e., performance of the a posteriori analysis was conditioned upon the presence of statistically significant main treatment effects. The first hypothesis set tested was identical to the *main research* hypothesis set stated previously, except that it included all levels of the treatment variable of Test Result Error. If the null hypothesis could not be rejected, no further analysis would be performed.

Hypothesis Set 5

H_0: The group means of scores for the task performance measures False Alarm, Missed Bad, and Bonus Score, respectrively, were not significantly different under each of the four levels of the Test Result Error treatment, i.e., Level 1 (zero error), Level 2 (25% of the test results erroneous), Level 3 (50% of the "good" test results changed to read "bad"), and Level 4 (50% of the "bad" test results changed to read "good").

H_a: The means were significantly different.

The overall analysis was also used to examine the treatment effects of Subject-To-Experimenter Assignment and those of Problem Set Order Of Presentation.

Hypothesis Set 6

This hypothesis set is the same as Hypothesis Set 5 except that the comparisons of group means were made with respect to the four conditions of the treatment of Subject-To-Experimenter Assignment.

Hypothesis Set 7

This hypothesis set is the same as Hypothesis Set 5 and Hypothesis Set 6 except that the comparisons of group means were made with respect to the two conditions of the treatment of Problem Set Order Of Presentation.

A Posteriori Analysis

In the event that the ANOVA in the overall analysis indicated the existence of significant differences among group means of the scores for any of the three task performance measures (False Alarm, Missed Bad, and Bonus Score), compared respectively across the various conditions within each of the three experimental treatments (Test Result Error, Subject-To-Experimenter Assignment, and Problem Set Order Of Presentation, an a posteriori analysis was to be performed. The statistical procedure selected for use in that analysis was the Tukey HSD (Honestly Significant Difference) Test. That test is specifically designed for an a posteriori analysis of data to determine which of a series of treatment conditions account for significant differences among experimental treatments which are shown to be statistically significant in the results of an overall ANOVA. The Tukey HSD Test is more statistically powerful for that purpose than a series of comparisons using the Student's T-Test.

Hypothesis Set 8

H_0: No significant differences existed among the group means being compared, i.e., the group means associated with whatever (experimental treatment-by-task performance measure) main effect was shown to be significant by the ANOVA in the overall analysis.

H_a: The means were significantly different.

The null hypothesis was to be iteratively tested by a series of comparisons of successively smaller pairwise differences between group means with the critical value for the Tukey HSD (Honestly Significant Difference) Test. The series of tests was to be concluded when the largest remaining pairwise difference between group means did not exceed that critical value.

STATISTICAL ASSUMPTIONS

Subjects

1. Subjects were drawn from a normally distributed population, i.e., experimental errors for each treatment population were normally distributed and independent.

2. The variance in subjects' responses due to experimental error was homogeneous within each treatment population.

3. A subject's response was the sum of the effects denoted in the function which describes the linear model which underlies the experimental design employed.

Statistical Model

The choice of a model in the performance of the ANOVA is dependent on whether the treatment conditions are assumed to be fixed or random. In this experiment, all treatment conditions were assumed to be random. In this sense, random means that the treatment levels included in the experiment were a random sample from a much larger population of treatment levels. Thus, results of the experiment can be more readily generalized to that larger population than if the selection of treatment levels were to have been constrained for some reason or "fixed."

CHAPTER III

RESULTS AND FINDINGS

INTRODUCTION

It will be recalled, from Chapter I, that the primary research question was:

> In the presence of noise, is there a human operator bias present in man-machine based troubleshooting tasks which has a systematic affect on performance as measured by the commission of Type I (False Alarm) and Type II (Missed Bad) errors?

The second research question was:

> If the existence of a systematic human operator bias in man-machine based troubleshooting tasks can be demonstrated, what is the direction of its impact, i.e., does the bias selectively inhibit or facilitate task performance under conditions of test result uncertainty which favor either the commission of a Type I or a Type II error?

In Chapter II, the research questions were stated in terms of several statistically testable hypotheses and a methodology was described for operationally defining the research questions in terms of an experiment to gather data to answer them. Chapter II also contained a description of a three level analysis which was performed using the experimental data.

This chapter presents the results of that analysis. Results are formatted in terms of the three levels of the data analysis; (1) a priori orthogonal comparisons which directly address the primary research question, (2) an overall evaluation of the statistical significance of main experimental treatment effects on three measures of experimental task performance (False Alarm, Missed Bad, and Bonus Score), and (3) a posteriori paired comparisons of task performance measure mean scores to identify individual treatment conditions which make a statistically significant contribution to the main treatment effects that were determined to be statistically significant.

Results of the experiment will now be presented under chapter subheadings corresponding to the level of the data analysis in which they were obtained. The hypotheses referred to in the tables which follow were stated in Chapter II. They will not be reproduced in this chapter.

RESULTS OF THE STATISTICAL ANALYSIS

A Priori Analysis

Table 3 summarizes the results of the a priori analysis.

As can be seen from Table 3, the experimental treatment condition of fifty percent of the "good" test results changed to read "bad" did not have an effect (on either the False Alarm or the Missed Bad task performance measure scores) which was statistically different from that

TABLE 3

SUMMARY OF
A PRIORI ANALYSIS RESULTS

HYPOTHESIS SET	TASK PERFORMANCE VARIABLE	TEST RESULT ERROR LEVEL COMPARISON	OUTCOME OF COMPARISON (at $\alpha \leq .05$)
#1	FALSE ALARM	50% "good" called "bad" with 50% "bad" called "good"	Fail to reject Ho
#2	MISSED BAD	50% "good" called "bad" with 50% "bad" called "good"	Fail to reject Ho
#3	FALSE ALARM	25% error with 50% "good" called "bad" plus 50% "bad" called "good"	Fail to reject Ho
#4	MISSED BAD	25% error with 50% "good" called "bad" plus 50% "bad" called "good"	Fail to reject Ho

of the experimental treatment condition of fifty percent of the "bad" test results changed to read "good."

Table 3 also reveals that the experimental treatment condition of twenty-five percent of all test results changed to be erroneous did not have an effect (on either the False Alarm or the Missed Bad task performance measure scores) which was statistically different from that of the experimental treatment conditions of fifty percent of the "good" test results changed to read "bad," or of fifty percent of the "bad" test results changed to read "good."

Thus, the null forms of Hypothesis sets one, two, three, and four cannot be rejected, i.e, the null form of the primary research hypothesis cannot be rejected. Furthermore, failure to reject the null form of the primary research hypothesis rendered the second research question moot. No attempt was made to answer it.

Figures 4, 5, and 6 are plots of subject scores for the experimental task performance measures of False Alarm, Missed Bad, and Bonus Score, respectively, as a function of the experimental treatment of Test Result Error.

Overall Analysis

Table 4 presents a summary of the main (experimental treatment-by-experimental task performance measure) effects which were found to be statistically significant in an overall analysis of the entire set of experimental data. It also presents a summary of the (experimental treatment-by-experimental task performance measure) first order interactions which were found to be statistically significant. The statistical significance of effects was determined by use of the Analysis of Variance (ANOVA) procedure and the F statistic.

FIGURE 4

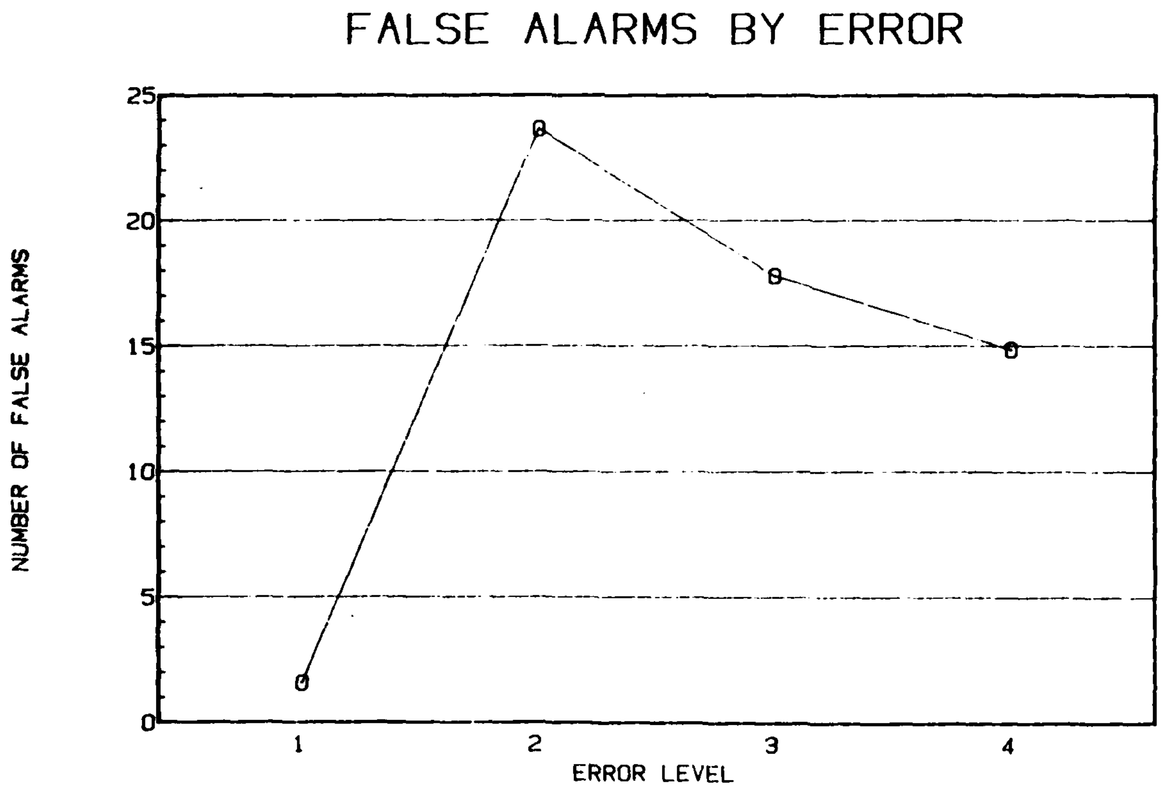

FALSE ALARMS BY ERROR
NUMBER OF FALSE ALARMS
25
20
15
10
5
0
1
2
3
4
ERROR LEVEL

FIGURE 5

MISSED BADS BY ERROR LEVEL

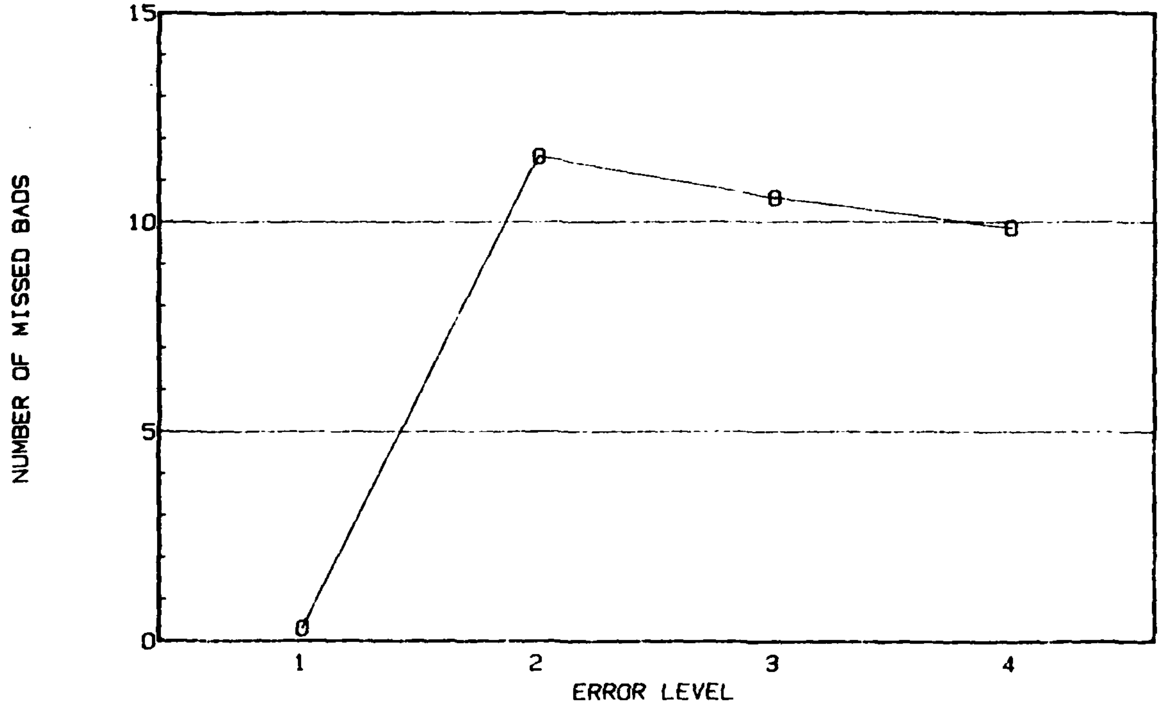
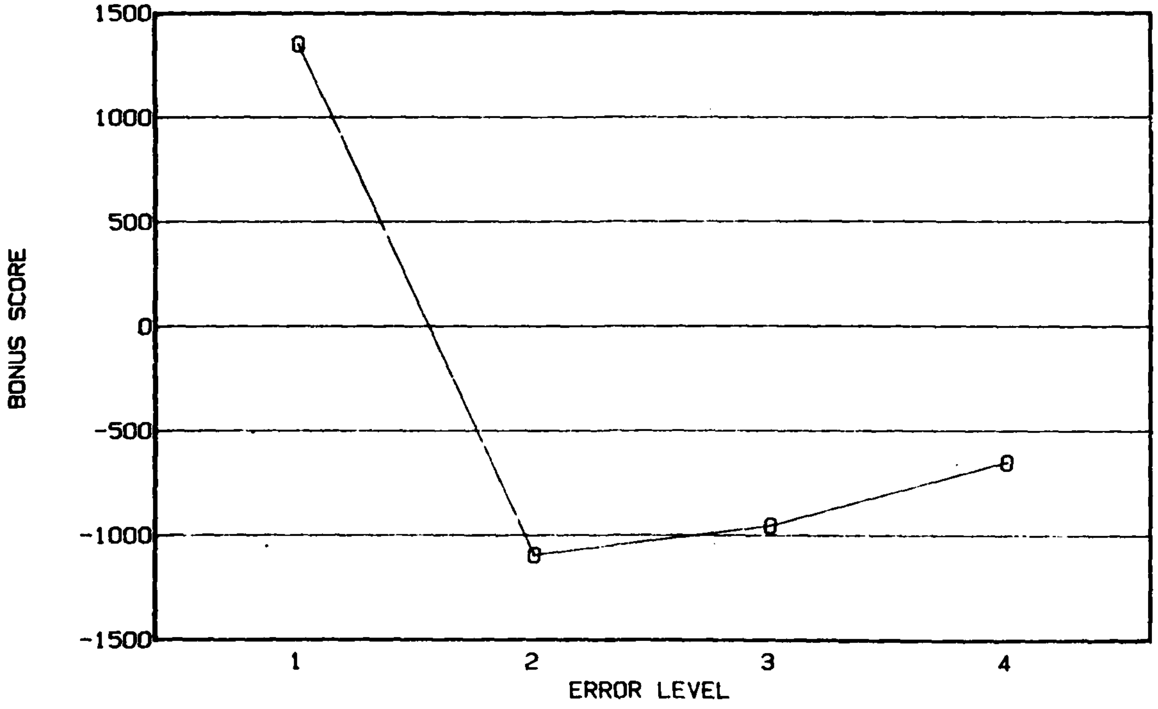


FIGURE 6

BONUS SCORE BY ERROR



As can be seen in Table 4, three main effects and two first order interaction effects[5] were shown to be statistically significant. However, the only experimental treatment which had a statistically significant effect on subjects' experimental task performance scores was Test Result Error. The experimental treatments of Subject-To-Experimenter Assignment and Problem Set Order Of Presentation did not produce a statistically significant effect on subjects' scores for any of the three measures of experimental task performance.

TABLE 4

SIGNIFICANT MAIN AND INTERACTION EFFECTS FROM THE ANOVA

TASK PERFORMANCE VARIABLE	TREATMENT MAIN EFFECT	LEVEL OF SIGNIFI-CANCE	TREATMENT INTERACTION EFFECT	LEVEL OF SIGNIFI-CANCE
FALSE ALARM	ERROR	$\alpha \leq .01$		
MISSED BAD	ERROR	$\alpha \leq .01$	ERROR BY SUB-TO-EXP ASSIGNMENT	$\alpha \leq .01$
BONUS SCORE	ERROR	$\alpha \leq .01$	ERROR BY SUB-TO-EXP ASSIGNMENT	$\alpha \leq .01$

[5] The implication of significant main effects, in the presence of significant interaction terms, is that the main effects may be significant only for certain levels of the independent variables of the ANOVA.

Results of the ANOVA are provided in greater detail in Tables 8, 9, and 10, which are contained in Appendix C.

Thus, the null form of Hypothesis Set five can be rejected with respect to the experimental task performance measures of False Alarm, Missed Bad, and Bonus Score. The null forms of Hypothesis Sets six and seven cannot be rejected with respect to any of three experimental task performance measures.

Table 5 is a summary of the results of the Overall Analysis. It provides a map of the (experimental treatment-by-experimental task performance measure) potential effects

TABLE 5

SUMMARY OF
OVERALL ANALYSIS RESULTS

HYPOTHESIS SET	TASK PERFORMANCE VARIABLE	EXPERIMENTAL TREATMENT	OUTCOME (at $\alpha \leq .05$)
#5	FALSE ALARM	TEST RESULT ERROR (ALL LEVELS)	Reject Ho
	MISSED BAD		Reject Ho
	BONUS SCORE		Reject Ho
#6	FALSE ALARM	SUBJECT-TO-EXPERIMENTER ASSIGNMENT (ALL CONDITIONS)	Fail to Reject Ho
	MISSED BAD		Fail to Reject Ho
	BONUS SCORE		Fail to Reject Ho
#7	FALSE ALARM	PROBLEM SET ORDER OF PRESENTATION (ALL CONDITIONS)	Fail to Reject Ho
	MISSED BAD		Fail to Reject Ho
	BONUS SCORE		Fail to Reject Ho

and indicates which of them were found to be statistically significant.

A POSTERIORI ANALYSIS

The a posteriori analysis examined the (experimental treatment-by-experimental task performance measure) main effects which were identified, in the Overall Analysis, to be statistically significant. It identified specific treatment conditions to which the statistical significance of main treatment effects could be attributed. Table 6 provides a summary of the results. Table 11, located in Appendix C, provides a more detailed presentation of the results.

TABLE 6

SUMMARY OF
A POSTERIORI ANALYSIS RESULTS

HYPOTHESIS SET	TASK PERFORMANCE VARIABLE	EXPERIMENTAL TREATMENT	OUTCOME (AT $\alpha \leq .05$)
#8	FALSE ALARM	TEST RESULT ERROR (ALL LEVELS)	Reject Ho for: Level 1 vs Level 2 Level 1 vs Level 3
#8	MISSED BAD	TEST RESULT ERROR (ALL LEVELS)	Reject Ho for: Level 1 vs Level 2 Level 1 vs Level 3 Level 1 vs Level 4
#8	BONUS SCORE	TEST RESULT (ALL LEVELS)	Reject Ho for: Level 1 vs Level 2 Level 1 vs Level 3 Level 1 vs Level 4

Results of the series of paired comparisons of group means for the four conditions (zero error, 25% error, 50% of the "good" test results changed to read "bad," and 50% of the "bad" test results changed to read "good") of the experimental treatment of Test Result Error were statistically significant only for comparisons which included the zero error treatment condition. Thus, the statistical significance of the Test Result Error treatment main effect found by the Overall Analysis could be attributed solely to the absence of test result error as compared to the presence of test result error.

As can be seen in Table 6, the above finding is true for each of the three experimental task performance measures. There is, however, one exception. The comparison of Level 1 and Level 4 (zero error vs 50% of the "bad" test results changed to read "good") of the Test Result Error treatment did not indicate a statistically significant difference between means of subjects' scores for the experimental task performance measure False Alarm.

Thus, the null form of Hypothesis Set 8 can be rejected only as is indicated in the Outcome Section of Table 6.

SUMMARY

Findings of the three level analysis of the experimental data are provided below in list form.

1. The null form of the primary research hypothesis could not be rejected at the five percent level of statistical significance, in a two-tailed test of significance.

2. The second research question was determined to be moot on the basis of the above finding.

3. Test result error appeared to influence subjects' performance of the experimental task only on the basis of its presence or absence.

CHAPTER IV

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

SUMMARY

A major problem facing the Air Force in the area of aircraft maintenance troubleshooting was described in Chapter I. That problem is the high incidence of errors committed by troubleshooting personnel such that either properly functioning equipment is removed from a system for repair or faulty equipment remains unnoticed and is left within a system.

Two research questions were developed which addressed the existence and directionality of human bias in the performance of troubleshooting tasks in the presence of noise. The premise of those questions was; if operator bias is a factor which systematically affects troubleshooting performance, it may be quantifiable and used to advantage in reducing the incidence of troubleshooting error. Given the similarity of many troubleshooting tasks to the task of signal detection in a background of noise, answers to the two research questions were thought to be useful in determining whether powerful signal detection-in-noise analytical tools such as the Relative Operating Characteristic (ROC) curve would be useful in reducing troubleshooting error.

The research questions were stated as statistically testable hypotheses and an experiment was devised to empirically test them. Chapter II provided a description of the research approach and the methodology used to gather data to answer the research questions. Chapter III presented the results and findings of the experiment.

CONCLUSIONS

Results of the experiment conducted as part of this thesis did not indicate the existence of human bias as a factor which significantly affects human performance in troubleshooting tasks. However, an analysis of the experimental data revealed one experimental treatment and task performance measure combination which provided empirical evidence that the introduction of at least one kind and amount of error into troubleshooting test results has little or no effect on troubleshooting task performance.[5] That finding and the fact that college students were used as

[5] The a posteriori analysis of the experimental data revealed that the combination of the experimental treatment condition of fifty percent of the "bad" test results changed to read "good" did not produce an effect on subject scores for the experimental task performance measure False Alarm which was statistically different from that produced by the experimental treatment condition of zero error introduced into test results.

subjects in the experiment, as opposed to actual maintenance troubleshooting personnel, provide grounds sufficient to call into question a conclusion that human bias does not exist as a factor which significantly affects human performance in troubleshooting tasks.

Only two conclusions could be reached on the basis of the experimental results:

1. The experiment did not provide empirical evidence sufficient to conclude that human bias is or that it is not a factor which significantly affects human performance in troubleshooting tasks.

2. The introduction of error into troubleshooting test results has a significant effect on troubleshooting task performance.

RECOMMENDATIONS

The lack of definitive results from the experiment conducted in this thesis precluded the setting forth of strong recommendations. However, lessons learned from its conduct suggest that further experimentation in this area should (1) emphasize the application of rigid controls on all aspects of experimental treatments, and (2) pay considerable attention to subject selection, training, and motivation for taking part in the experiment as potential sources of difficulty.

APPENDICES

APPENDIX A

INSTRUCTIONS GIVEN TO SUBJECTS

TRAINING SESSION INSTRUCTIONS

INTRODUCTION OF EXPERIMENT TO SUBJECTS

I. General Information

1. You are invited to participate in a study of how people resolve problems in electronic equipment. This is called troubleshooting. We hope to learn more about how humans perform in troubleshooting situations.

II. Specific Experimental Description

1. This task consists of solving two sets of thirty troubleshooting problems, for which you will receive training.

2. Your responses will be recorded for analysis at a later date. You will not be *identified in any way* with these data upon completion of your participation. However, these data, devoid of any means of identifying you, will be kept for future reference and possible additional analysis.

III. Subject Participation

1. After sufficient training, which takes about one hour for each of three practice sessions, you will be asked to participate in two experimental data collection sessions, each lasting about 50-60 minutes. You may terminate your participation at any time.

IV. Other Information

1. You will be provided a detailed briefing of the experimental procedure. This should aid you in understanding the research in which you are participating. If you have any questions, please ask them. The results of this study will be available to you upon its completion.

INITIAL TRAINING

This experiment consists of two sets of thirty problems based on information which will be presented to you in a display on the computer screen. The display depicts a simulated system of twenty-five electronic components which are represented by a network of boxes. The network shows how the boxes are connected with one another. Each box is numbered for easy identification. In each of the problems, you will be required to (1) determine if there is a malfunctioning (bad) box in the network, (2) to replace it if one is found, and (3) to assure yourself that the network contains only properly operating (good) boxes before going on to the next problem.

There is only one network to consider. It will remain the same for all of the problems. In addition, there will never be more than one bad box within any single problem.

The boxes within the network operate as signal processors, i.e., they receive a signal, do something to it, and pass the signal on to another box. All signals are passed on from the left to the right of the network. A good signal is always entered into the boxes at the far left. There are several boxes which are further to the right in the network which are not connected to any boxes to the left of them. They, also, always receive a good signal. A bad box transforms a good signal into a bad signal as it processes it. That bad signal is then passed on to boxes which are to the right of the bad box, thus making bad the output signals of all boxes which are to the right of the bad box and which are connected to it.

The signal coming out of a box is good if and only if:

1) all signals leading into that box are good,

AND

2) the box itself is also good.

Otherwise, the signal coming out of the box will be bad.

Four actions are available to you as you work on each problem. They provide the means for you to perform tests on the network, replace boxes, and designate the completion of a problem. They are:

1) an Overall System Test,

2) a bad box Localization Test,

3) a Replacement Command, and

4) a Problem Completion Command which will initiate the next problem.

1) Overall System Test - This action initiates a test of the complete network. Results are shown by five large numerals which will appear to the right of the column of boxes on the far right side of the network. They indicate the output from each of those boxes. It should be remembered that if the output of a box is good, it may be presumed that the outputs of all of the boxes which are to the left of it and which are connected to it are also good. The numerals which indicate the results of the test will be either a one or a zero. A one indicates a good output signal. A zero indicates a bad output signal. The test result numbers will remain visible on the display until the next action is taken.

2) Localization Test - This action initiates a test of either a single box or a series of boxes which are connected. The selection of which box or series of boxes is to be tested is up to you. If you select a series of boxes to be tested as a unit, remember that they must be connected. The results of the test will indicate if a good signal entering the first box (the one farthest to the

left) emerges as a good signal from the last box (the one farthest to the right) in the series which you selected. If there is a bad box in the series, it will pass on a bad signal to other boxes in the series which are to its right in the network. Unlike the Overall System Test which assesses the final system outputs, i.e., the outputs of each box in the column of boxes on the far right side of the network, this test will merely give a single result of "good" or "bad." If you select a single box to be tested, the test will reveal if that box is good or bad. If you select a series of connected boxes to test, the test will only reveal that either all boxes in the series are good or that there is a box in the series that is bad. The test will not specify which box is bad. You will notice that, for any series of boxes, there may be many paths which a signal can take in going from the first to the last box in the series. The selection of which boxes to test should be made with care because the testing of a larger series of boxes, while covering more ground in a single step, will not yield as much specific information as a test of a small series (unless, of course, all of the boxes in the series test out as being good). Clearly, the use of a testing strategy will result in a need for fewer tests and will improve your score. If you attempt to test a series of boxes which are not connected with each other, you will be informed that the test is not valid. The attempt will not count against your score for the problem. Such a test is unreasonable because signals cannot flow between boxes which are not connected.

3) Replacement Command - This action initiates the replacement of a box which you have decided is bad with a good box. Keep in mind when performing this action that it will count against your score if you replace a box that is really not bad. Also, after replacing a box, you should assure yourself that the entire network of boxes is operating properly before going on to the next problem.

4) Problem Completion Command - This action specifies that you are sure that the entire network of boxes is operating properly, the problem is completed, and that you are ready for the next problem. Once this action is taken, there is no going back for any more tests. Be certain that you have completed all tests which you wish to make on a problem before taking it.

The experimenter will now instruct you on the use of the computer keyboard until you are thoroughly familiar with the combination of keys to press to take the actions described above. He will also demonstrate how to correct any errors you may make in typing in your commands to the computer.

Feel free at this time to request additional explanation of the task you are to perform. We want you to understand it and feel comfortable with the use of a computer, screen, and keyboard. This and the next two sessions will be devoted to training you for that purpose. The actions you may take during the course of a problem session and the keyboard entries which you must make to take them will be demonstrated and explained until you are satisfied that you understand both them and the objective of the task.

The following keyboard entries will now be explained and demonstrated:

1. Selecting an Action
2. Typing in an Overall System Test Command
3. Typing in a Localization Test Command
4. Typing in a Replacement Command
5. Correcting Keyboard Entry Errors
   A. Localization Test
      (1) Backspace Keying
      (2) Specification of an Invalid Test
   B. Replacement
      (1) Backspace Keying
      (2) Specification of an Invalid Replacement

The experimenter will now work with you, step by step, through several problems and will aid you on several others until you are able to solve problems by yourself.

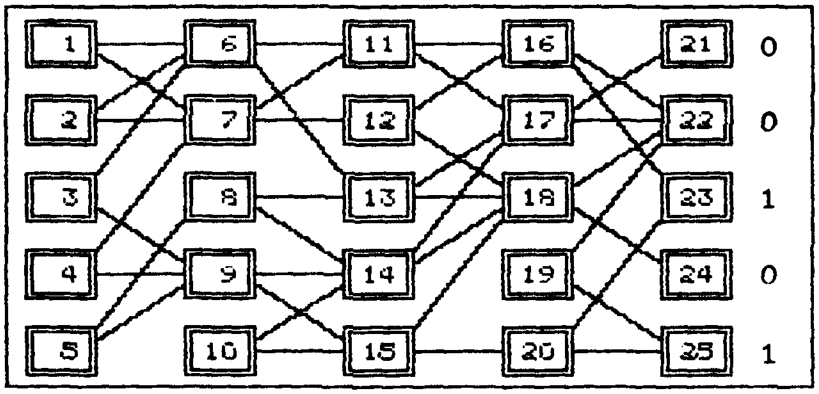
EXAMPLE TRAINING PROBLEM

The following problem[7] illustrates the training session regimen and the guidance afforded by the instructors to each subject. Care was taken to point out the generalizations which can be drawn from the examples in order to emphasize the importance of strategy formulation to efficiency in problem solving.

Questions were answered and encouraged to ensure that each subject became thoroughly familiar with the experimental task, the mechanics of performing it, and the experimental environment. Each of the four instructors took pains to provide an encouraging atmosphere aimed at eliminating any aspect of the experimental environment which might prove to be threatening to the subjects.

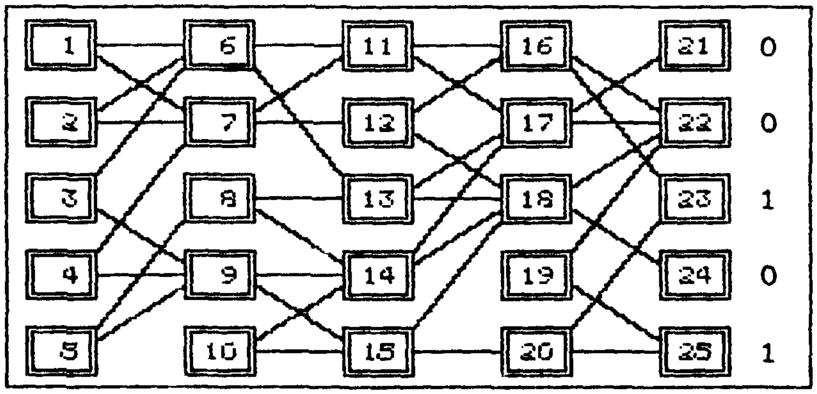
[7] The illustrative problem contained in this appendix is one of seven which were shown to subjects and solved by the experimenter prior to beginning the three training sessions.

PROBLEM #1 - Run the Overall Test



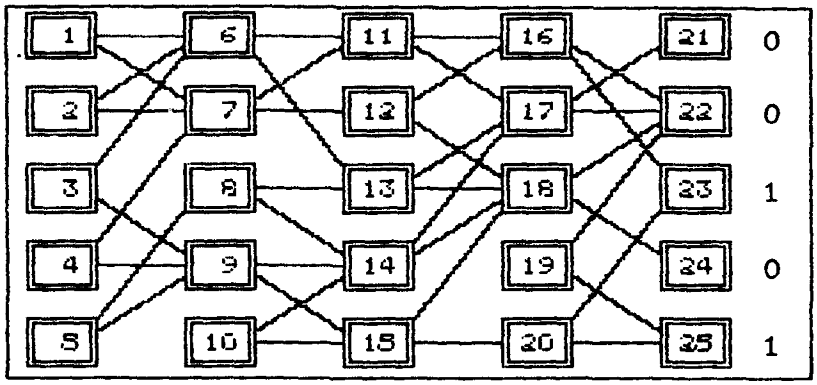
This is the display you will see when you run an overall test. The results are displayed at the far right side of the screen. A "one" beside a box in the far right column means that it and every box which is connected to it is good. A "zero" means that either that box or a box which is connected to it is bad. As you will see, strong inferences can be made from the overall test and the pattern of connections among the boxes.

PROBLEM #1 - *continued*



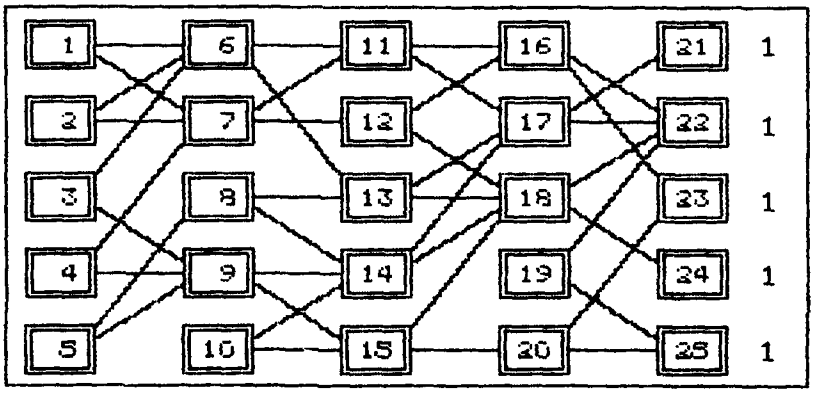
Noting that any box connected to a good box in the extreme right side column must be good, we can eliminate all but eight boxes which might be bad (8,13,14,17,18,21,22, and 24).

PROBLEM #1 - continued



Since only one box can be bad, boxes 17, 18, 21, 22, and 24 must be good because they could not produce the results of the test as shown. Therefore, the malfunction must be in either box 8 or 13 or 14. We can now go on to the localization test to find the bad box.

PROBLEM #1 - continued



Having reduced the problem this far, we must now run a localization test to find the bad box. We could test each box separately, or in pairs, or (providing that they are connected to each other) in any combination which forms a circuit. Some tests are more efficient than others. For this example, test 8 to 13. Since this reads good, box 14 must be the bad box. Replace it, run the overall test to check the results of the replacement, and move on to the next problem. The results of the final overall test are shown above.

END OF FINAL TRAINING SESSION

By now you have a good grasp of how to operate the computer in solving the problems that are presented. In fact, you are probably performing the tests in an almost theoretically "perfect fashion."

In order to make the experimental sessions a little more interesting than the training sessions, we are going to make two changes:

1) For some people, the tests will make mistakes and tell them that a test reads good when it really should say that the test reads bad, and/or vice versa. Both the Localization Test and the Overall Test will make these mistakes. Such mistakes will be made about 1/4 of the time.

2) We are going to offer you the chance to earn up to $2.00 per hour more than the going rate of $3.55 per hour for your participation. You can do this by performing very well on the two experimental problem sets. The program will keep track of your actions and give you 100 points for every bad part that you replace. However, you lose points for other actions as follows:

-2 points for every localize test

-10 points for every removal

-100 points to leave a bad part in

The overall test is free. You neither lose nor gain points by making it. You should also note that the level of error may not be the same for each subject. To make it fairer for those who have different levels of error in their problems, we will only compare your score with others who have the same level of error. The top person in each group gets $2.00 extra per hour; the low person gets nothing extra. You will be told your score after your last session.

We also wish to remind you that you can drop out of the experiment at any time. Just get up and go! You will be remunerated for your time up to that point, and receive commensurate credit toward your course, if you are getting any. If you cannot attend a scheduled session, we will reschedule you for another time. Please give us as much notice as possible.

Be sure to take note of when you are scheduled to participate in the experimental sessions, and try to arrive on time. Thank you.

APPENDIX B

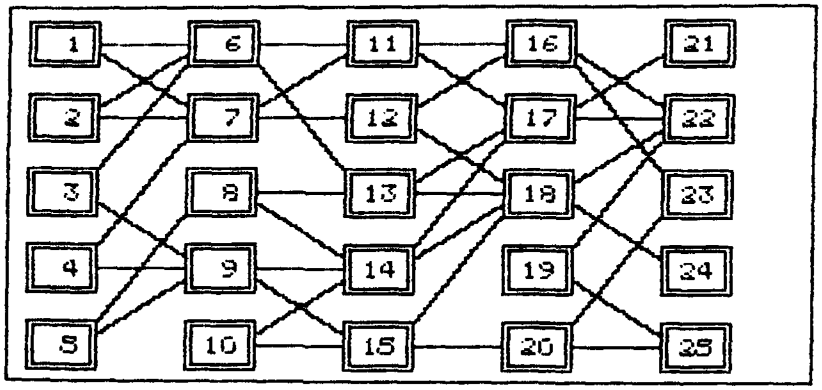
DESCRIPTION OF EXPERIMENTAL SUPPORT SOFTWARE

Several computer programs were developed to generate experimental displays, provide for a discourse between subjects and the computer during training and experimental sessions, create and maintain files for recording and processing subjects' responses and other data concerning the logistics of running each of sixty-four subjects through three training and two experimental sessions. Short descriptions of them are provided below.

<u>TASK GENERATOR</u> - This program served the dual purpose of generating and drawing the network used as the experimental problem network and of generating the problem sets. The network information was stored as two files. The first of these was the NETWORK DATA file which was a twenty by two matrix listing the two network nodes (boxes) to which the first twenty network nodes were connected. Figure 7 illustrates the fact that, moving from left to right, each of the first twenty boxes has only two direct connections to boxes which are further to its right in the network. Those connections are to boxes, in the next column of boxes, to the right of the subject box.

FIGURE 7

EXPERIMENTAL NETWORK



The NETWORK DATA file was read during the training and experimental sessions to create another matrix to be used in running the troubleshooting tests of the network which were to be made by subjects as part of the experimental task. The second file in which network information was stored was a memory dump of the two thousand (hexadecimal) core storage locations containing the information which constituted the graphic representation of the network. This file was created to obviate the need to have the computer redraw the picture of the network on the CRT for each session. The file allowed the network picture to be reconstituted instantaneously, as a whole, under the control of the CONFIGURATION program which will be described below. The five problem sets generated by the TASK GENERATOR program were also stored as data files to be called up later as needed by the CONFIGURATION program. They were generated on a random basis such that approximately fifty percent of the problems in each problem set contained no bad boxes, while the remaining fifty percent of the problems contained a bad box which was randomly selected from the twenty five nodes (boxes) in the network.

CONFIGURATION - This program served the purpose of dealing with the CONTROL file in automatically sequencing subjects through the experiment in the correct order of problem set presentation. The CONTROL file for a particular subject is written by the program at the beginning of a subject's first training session and contains the experimenter identification, the subject identification, the next session number, the error condition of the problem set, and information concerning the planned order of presentation of problem sets. In subsequent sessions, the CONFIGURATION program was loaded and ran the main experimental program. It also updated the CONTROL file. The purpose of the CONFIGURATION program was to allow the experimenters to minimize their interaction with subjects during the experimental sessions. It allowed an experimenter to remove himself from the experimental environment after merely identifying himself and the subject to the computer at the beginning of the first session. The experimenter was blind to the main independent variables of the experiment, i.e., the test result error condition of the problem set, the order of problem set presentation to the subject, and the subject-to-experimenter assignment.

EXPERIMENT II - This program was the only program that a subject dealt with during the experiment. It generated the experimental tasks. Controlled by the CONFIGURATION program, it first loaded the file containing the graphic representation of the network and then read the current values in the control file. It called up the appropriate problem set, for the session and subject specified by the CONTROL file, and initiated the experimental session with the subject. As subject response data were generated, the program recorded them on a scratch file and later, upon a subject's designation of problem completion, copied the contents of the scratch file to a permanent subject response data file. This manner of recording the data allowed for the correction of data entry errors the subject might make, before they become a part of the permanent data record. Allowable subject responses which were recorded are; (1) Overall Test, (2) Localization Test, (3) Replacement of a Box, and (4) End of Problem Designation. This method of recording subject responses ensured that all responses were immediately totaled, kept separate, and formatted in a way which facilitated the retrieval of data to support the performance of several kinds of computerized analyses. At the end of each session, the experimenter, subject, and session designations were added to each data record.

FIX - This program contains several utility routines for use during the course of the conduct of the experiment. These were used to (1) re-set the CONTROL file so that a new subject could be run on a data recording disk, (2) read out an experimental session data file, (3) tally the number of the various actions performed by a subject during individual sessions, (4) catalog data files, and (5) concatenate data files. Each of these five functions is accomplished by separate routines either within or called up by the FIX program as required to fulfill a user command.

TALLY - This program is called up by the FIX program to function within it in order to tally the subject responses by session. The outputs are (1) quantity of Overall Tests, (2) quantity of Localization Tests, (3) quantity of Replacement of Box actions, (4) quantity of correct replacement actions, (5) quantity of False Alarms (incorrect replacement of a good box), (6) quantity of Missed Bad (failures to locate the bad box in a problem, if there was one), and (7) the Bonus Score earned by a subject (dependent on a series of values placed on the quantity and correctness of various possible subject responses).

APPENDIX C

STATISTICAL ANALYSIS SUMMARY TABLES

TABLE 7

EQUATION FOR STATISTICAL MODEL AND EXPECTED MEAN SQUARES FOR ERROR TERMS

STRUCTURAL MODEL

$$X_{ijkm} = M + A_i + G_k + AG_{ik} + L_{m(ik)} + B_j + AB_{ij} + BG_{jk} + ABG_{ijk} + BL_{jm(ik)} + R_{o(ijkm)}$$

N	SOURCE	E(MS)	ERROR TERM
z=4	T	$Y_e^2 + qY_p^2 + nqrY_a^2$	S(TE)
q=4	E	$Y_e^2 + qY_p^2 + nzqY_g^2$	S(TE)
	TE	$Y_e^2 + qY_p^2 + nqY_{ag}^2$	S(TE)
n=4	S(TE)	$Y_e^2 + qY_p^2$	
r=2	P	$Y_e^2 + Y_{bp}^2 + nzrY_b^2$	BS(TE)
	TP	$Y_e^2 + Y_{bp}^2 + nrY_{ab}^2$	BS(TE)
	PE	$Y_e^2 + Y_{bp}^2 + nzY_{bg}^2$	BS(TE)
	TPE	$Y_e^2 + Y_{bp}^2 + nY_{abg}^2$	BS(TE)
	BS(TE)	$Y_e^2 + Y_{bp}^2$	

TABLE 8

ANOVA SUMMARY FOR
STANDARDIZED FALSE ALARM
TASK PERFORMANCE MEASURE

SV	SS	df	MS	error term	F
1. ERROR	33.149	3	11.050	3+6-8 (df=11)	7.038**
2. TRAINER	2.548	3	.849	4+8-9 (df=41)	0.569
3. E*T	11.947	9	1.327	4+8-9 (df=41)	0.889
4. S(ET)	62.453	42	1.487	9	17.702**
5. PROB SET	.028	1	.028	8	0.311
6. E*P	.999	3	.333	8	3.7
7. P*T	.233	3	.078	8	0.867
8. E*P*T	.809	9	.090	9	1.071
9. P*S(ET)	3.545	42	.084	--	--
TOTAL		115			

* P ≤ .05   ** P ≤ .01

TABLE 9

ANOVA SUMMARY OF
STANDARDIZED MISSED BAD
TASK PERFORMANCE MEASURE

SV	SS	df	MS	error term	F
1. ERROR	72.791	3	24.264	3+6-8 (df=10)	14.590**
2. TRAINER	2.848	3	.949	4+8-9 (df=33)	2.090
3. E*T	13.799	9	1.553	4+8-9 (df=33)	3.377**
4. S(ET)	18.761	42	.447	9	4.382**
5. PROB SET	.012	1	.012	8	0.110
6. E*P	.718	3	.239	8	2.193
7. P*T	.567	3	.189	8	1.734
8. E*P*T	.977	9	.109	9	1.069
9. P*S(ET)	4.292	42	.102	--	--
TOTAL		115			

* P≤.05 ** P≤.01

TABLE 10

ANOVA SUMMARY FOR
STANDARDIZED BONUS SCORE
TASK PERFORMANCE MEASURE

SV	SS	df	MS	error term	F
1. ERROR	84.778	3	28.259	3+6-8 (df=9)	25.459**
2. TRAINER	1.392	3	.464	4+8-9 (df=27)	1.381
3. E*T	9.470	9	1.052	4+8-9 (df=27)	3.131**
4. S(ET)	13.725	42	.327	9	3.175**
5. PROB SET	.007	1	.007	8	0.0625
6. E*P	.511	3	.170	8	1.518
7. P*T	.725	3	.242	8	2.161
8. E*P*T	1.004	9	.112	9	1.087
9. P*S(ET)	4.315	42	.103	--	--
TOTAL		115			

* P≤.05 ** P≤.01

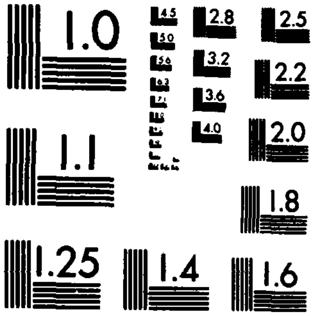
TABLE 11

A POSTERIORI COMPARISONS OF
PAIRED GROUP MEANS
FOR THE TREATMENT EFFECT OF
TEST RESULT ERROR

	LEVEL 1 (L1) ZERO ERROR GROUP MEAN	LEVEL 2 (L2) 25% ERROR GROUP MEAN	LEVEL 3 (L3) 50% OF "GOOD" CALLED "BAD" GROUP MEAN	LEVEL 4 (L4) 50% OF "BAD" CALLED "GOOD" GROUP MEAN	DEGREES OF FREEDOM	K	MEAN SQUARE ERROR	CRITICAL VALUE FOR TUKEY HSD TEST	SIGNIFICANT OUTCOME ($\alpha \leq .05$)
FALSE ALARM	-.8317	.4764	.4832	-.1108	11	4	1.57	1.2755	L1 vs L2 L1 vs L3
MISSED BAD	-1.3093	-.6713	-.4687	-.3465	6	4	1.663	1.6315	L1 vs L2 L1 vs L3 L1 vs L4
BONUS SCORE	1.4151	-.7135	-.5669	-.3002	9	4	1.110	1.1996	L1 vs L2 L1 vs L3 L1 vs L4

AD-A128 584 EFFECT OF TEST RESULT UNCERTAINTY ON THE PERFORMANCE OF 2/2
A CONTEXT-FREE TR..(U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH SCHOOL OF SYST.. H A BARAN
UNCLASSIFIED 17 DEC 82 AFIT-LSSR-86-82 F/G 15/5 NL

END
DATE
FILMED
[illegible]
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SELECTED BIBLIOGRAPHY

A. REFERENCES CITED

Clyman, M., Grenetz, P. S., and R. S. Schultz, The Economics of Maintenance Improvement Feasibility Study (Phase I). Interim Report No. ISI-W-7858-8, Information Spectrum, Inc., 955 Louis Drive, Warminster PA, April 1978.

Comptroller General of the United States, Effectiveness of U. S. Forces Can Be Increased Through Improved Weapon System Design. Report to the Congress PSAD-81-17, United States Government Accounting Office, Washington DC, January 1981.

Egan, J. P., Signal Detection Theory and ROC Analysis. New York: Academic Press, 1975.

Gibson, Major P., USAF, "Automated Diagnostic Systems." A briefing given at the AFHRL/RADC Diagnostic Errors Workshop, Wright-Patterson AFB OH, March 1982.

Gold, D., Kleine, B., Fuchs, F., Ravo, S., and K. Inaba, Aircraft Maintenance Effectiveness Simulation (AMES) Model; Final Report. Technical Report NAVTRAEQUIPCEN 77-D-0028-1, Naval Training Equipment Center, Orlando FL, 1980.

Green, D. M., and J. A. Swets, Signal Detection Theory and Psychophysics. New York: John Wiley and Sons, 1966.

Henneman, R.L., Measures of Human Performance in Fault Diagnosis Tasks. MSIE Thesis, Report T.107, University of Illinois, Urbana-Champaign IL, June 1981.

Henneman, R.L., and W.B. Rouse, Measures of Human Performance in Fault Diagnosis Tasks. Unpublished research report, 1982.

Hunt, R.M., and W.B. Rouse, "Problem Solving Skills of Maintenance Trainees in Diagnosing Faults in Simulated Powerplants," Human Factors, Vol. 23, No. 3, pp. 317-328, 1981.

IDA, Built-In-Test Equipment Requirements Workshop. IDA Paper P-1600, Program Analysis Division, Institute for Defense Analyses, Arlington VA, August 1981.

Johnson, W.B., and W.B. Rouse, "Computer Simulations for Fault Diagnosis Training: From Simulation to Live System Performance," Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles CA, October 1980.

________, "Training Maintenance Technicians for Troubleshooting: Two Experiments with Computer Simulations," Human Factors, Vol. 24, No. 3, pp. 271-276, 1982.

Kerr, T. H., "Failure Detection Aids for Human Operator Decisions in a Precision Inertial Navigation System Complex," Proceedings of the Symposium on Applications of Decision Theory to Problems of Diagnosis and Repair, Fairborn OH, AD A032205, 1976.

Kirk, R. E., Experimental Design: Procedures for the Behavioral Sciences. Belmont CA: Brooks/Cole Publishing Company, 1968.

Lipa, J.F., "Causes of Unnecessary Removals." A briefing given at the AFHRL/RADC Diagnostic Errors Workshop, Wright-Patterson AFB OH, March 1982.

Lomov, B.F., "The Analysis of the Operator's Activities in the Man-Machine System," Ergonomics, Vol. 22, No. 6, pp. 613-619, 1979.

McNicol, D., A Primer of Signal Detection Theory. London: George Allen & Unwin, 1972.

Neyman, J., and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," Philosophy Transactions Royal Society, London, Series A, p. 289, 1933.

Orlansky, J., and J. String, The Performance of Maintenance Technicians on the Job. IDA Paper P-1597, Science and Technology Division, Institute for Defense Analyses, Arlington VA, August 1981.

Owens, P. R., St. John, M. R., and F. D. Lamb, Avionics Maintenance Study. AFAL-TR-77-90, Air Force Avionics Laboratory, Wright-Patterson AFB OH, 1976.

Perry, R., Comparisons of Soviet and U.S. Technology. R-827-PR, The Rand Corporation, Santa Monica CA, June 1973.

________, The Interaction of Technology and Doctrine in the USAF. P-6281, The Rand Corporation, Santa Monica CA, January 1979.

Peterson, W. W., Birdsall, T. G., and W. C. Fox, "The Theory of Signal Detectability," Transactions of the IRE Professional Group on Information Theory, PGIT-4, pp. 171-212, 1954.

Pieper, W.J. and S. D. Folley, Effect of Ambiguous Test Results on Troubleshooting Performance. AMRL-TR-67-160, Aerospace Medical Research Laboratories, Wright-Patterson AFB OH, November 1967.

Rasmussen, J., and W.B. Rouse, Human Detection and Diagnosis of System Failures. New York: Plenum Press, 1981.

Rouse, W.B., "Human Problem Solving Performance in a Fault Diagnosis Task," IEEE Transactions on Systems, Man, and Cybernetics, SMC-8, No. 4, pp. 258-271, April 1978.

________, "Problem Solving Performance of Maintenance Trainees in a Fault Diagnosis Task," Human Factors, Vol. 21, No. 2, pp. 195-203, April 1979. (a)

________, "A Model of Human Decision Making in Fault Diagnosis Tasks that Include Feedback and Redundancy," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-9, No. 4, pp. 237-241, April 1979. (b)

Rouse, W.B., and R.M. Hunt, "Human Problem Solving in Fault Diagnosis Tasks." Report No. 82-2, Center for Man-Machine Systems Research, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta GA, July 1982.

Swets, J. A., "Is There a Sensory Threshold?" Science, Vol. 134, pp. 168-177, 1961.

Swets, J. A., "The Relative Operating Characteristic in Psychology," Science, Vol. 182, pp. 990-1000, 1973.

Swets, J. A., and R. M. Pickett, Evaluation of Diagnostic Errors, Unnumbered Manuscript, Bolt Beranek and Newman, Inc., Cambridge MA, 1980.

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., and B. A. Freeman, "Assessment of Diagnostic Technologies," Science, Vol. 205, pp. 753-759, 1979.

Wald, A., Statistical Decision Functions. New York: John Wiley and Sons, 1950.

B. RELATED SOURCES

Baran, H.A., and J.C. Goclowski, Digital Avionics Information System (DAIS): Life Cycle Cost Impact Modeling System (LCCIM) - A Managerial Overview. AFHRL-TR-79-64, Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson AFB OH, November 1980.

Bond, N.A., Jr. and J.W. Rigney, "Bayesian Aspects of Troubleshooting Behavior," Human Factors, Vol. 8, pp. 377-383, 1966.

Brooke, J.B., and K.D. Duncan, "Effects of System Display Format on Performance in a Fault Location Task," Ergonomics, Vol. 24, No. 3, pp. 175-189, 1981.

Brooke, J.B., Duncan, K.D., and E.C. Marshall, "Interactive Instruction in Solving Fault Finding Problems," International Journal of Man-Machine Studies, Vol. 10, pp. 603-611, 1978.

Brown, J.S., Burton, R.R., and A.G. Bell, "SOPHIE: A Step Toward Creating a Reactive Learning Environment," International Journal of Man-Machine Studies, Vol. 7, pp. 675-696, 1975.

Crawford, A.M., and K.S. Crawford, "Simulation of Operational Equipment With a Computer-based Instructional System: A Low Cost Training Technology," Human Factors, Vol. 20, pp. 215-224, 1978.

Crooks, W.H., Kuppin, M.A., and A. Freedy, Application of Adaptive Decision Aiding Systems to Computer Assisted Instruction: Adaptive Computerized Training System. Technical Report No. PATR-1028-77-1, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington VA, January 1977.

Dale, H.C.A., "Fault-Finding in Electronic Equipment," Ergonomics, Vol. 1, pp. 356-385, 1957.

Fink, C.D., and E.L. Shriver, Simulations for Maintenance Training: Some Issues, Problems, and Areas for Future Research. AFHRL-TR-78-27, Technical Training Division, Air Force Human Resources Laboratory, Lowry AFB CO, July 1978.

Glaser, R., Damrin, D.E., and F.M. Gardner, "The Tab Item: A Technique for the Measurement of Proficiency in Diagnostic Problem Solving Tasks," *Educational and Psychological Measurement*, Vol. 14, pp. 283-293, 1954.

Glass, A.A., *Problem-solving Techniques and Troubleshooting Simulators in Training Electronic Repairmen.* Unpublished Doctoral Thesis, Columbia University, New York, 1967.

Goldbeck, R.A., Bernstein, B.B., Hillix, W.A., and M.H. Marx, "Application of the Half-split Technique to Problem-solving Tasks," *Journal of Experimental Psychology*, Vol. 53, pp. 330-338, 1957.

Hunt, R.M., *A Study of Transfer of Problem Solving Skills from Context-Free to Context-Specific Fault Diagnosis Tasks.* Unpublished Master's Thesis, University of Illinois, Urbana IL, 1979.

Johannsen, G., and W.B. Rouse, "Mathematical Concepts for Modeling Human Behavior in Complex Man-Machine Systems," *Human Factors*, Vol. 21, No. 6, pp. 733-747, 1979.

Johnson, W.B., Rouse, S.H., and W.B. Rouse, *An Annotated Selected Bibliography on Human Performance in Fault Diagnosis Tasks.* Report No. TR-435, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria VA, January 1980.

King, W., "New Concepts in Maintenance Training," *Aviation Engineering and Maintenance*, Vol. 6, pp. 24-26, 1978.

Kleinman, D.L., Baron, S., and W.H. Levison, "*An Optimal Control Model of Human Response. Part I: Theory and Validation*," Automatica, Vol. 6, pp. 357-369, 1970.

Mallory, W.J., and T.K. Elliot, *Measuring Troubleshooting Skills Using Hardware-free Simulation.* AFHRL-TR-78-47, Technical Training Division, Air Force Human Resources Laboratory, Lowry AFB CO, 1978.

Miller, R.B., Folley, J.D., and P.R. Smith, Systematic Troubleshooting and the Half-Split Technique. Air Force Research and Development Command Technical Report 53-21, Human Resources Center, Air Force Research Center, Bolling AFB, Washington 25, DC, 1953.

Rasmussen, J., and A. Jensen, "Mental Procedures in Real-Life Tasks": A Case of Electronic Trouble Shooting," Ergonomics, Vol. 17, No. 3, pp. 293-307, 1974.

Rouse, W. B., "A Model of Human Decision Making in a Fault Diagnosis Task," IEEE Transactions on Systems, Man, and Cybernetics, SMC-8, No. 5, pp. 357-361, May 1978.

________, "Problem Solving Performance of First Semester Maintenance Trainess in Two Fault Diagnosis Tasks," Human Factors, Vol. 21, No. 5, pp. 611-618, 1979.

Rouse, W.B., and S.H. Rouse, "Measures of Complexity of Fault Diagnosis Tasks," IEEE Transactions on Systems, Man, Cybernetics, Vol. SMC-9, pp. 720-727, 1979.

Rouse, W.B., Rouse, S.H., and S.J. Pellegrino, "A Rule-Based Model of Human Problem Solving Performance in Fault Diagnosis Tasks," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-10, No. 7, 1980, pp. 366-376.

Steinman, J.H., Comparison of Performance on Analagous Simulated and Actual Troubleshooting Tasks. Memorandum SRM 67-1, U.S. Naval Personnel Research Activity, San Diego CA, July 1966.

Stolurow, L.M., Bergum, B., Hodgson, T. and J. Silva. "The Efficient Course of Troubleshooting as a Joint Function of Probability and Cost," Educational and Psychological Measurement, Vol. 15, No. 4, pp. 462-477, 1955.